Inter-Uni Datathon Presentation

Victorias_Most_Bitter

The Case:

- Recently, the company Skyline Financial Services (SFS), which is a leading digital bank that expertly runs data-driven insights, experienced a significant and unexpected dip in monthly revenue.
- An internal investigation revealed a pattern of fraudulent activities taking activity on our systems and slipping past existing detection systems.
- Our team, Victorias_most_bitter, was tasked with investigating this issue deeper to unveil the root cause of this fraudulent activity.

The development of the model:

1. Data cleaning: handled missing values, correcting inconsistencies, and removing duplicates.
2. Data exploration: Gaining insights about user demographics and behaviours to try identifying fraudulent behaviour.
3. Feature training and model training: picking features that correlate with fraudulent behaviour and applying statistical techniques to create predictive models.
4. Fine tuning and training: Tuning the model and testing for accuracy.

TransactionNumber	UserID	Age	Gender	Occupation	EducationLevel	MaritalStatus	NumDependents	Income	Expenditure	GiftsTransaction	TransactionDate	TransactionTim
8765	70	37	Female	Professional	Bachelor	Widowed	3	28884.43	14610.61	1050.01	2023-03-12	12:25:57
9645	3386	34	Male	Student	High School	Married	4	54919.07	39169.49	4969.71	2023-03-05	18:27:24
1145	2971	25	Male	Unemployed	Master	Married	2	74728.57	55873.76	1149.85	2023-11-10	17:16:56
15308	2925	25	Male	Professional	High School	Married	3	55712.62	89649.04	4335.7	2023-10-07	00:34:17
14967	2339	38	Male	Professional	High School	Single	4	53004.7	43601.02	4763.48	2023-09-22	18:40:08
15336	234	32	Female	Student	Bachelor	Single	1	101381.56	81036.3	196.29	2023-06-30	03:59:57
17281	1617	18	Male	Professional	High School	Single	1	63035.98	91610.14	3009.44	2023-11-21	15:00:38
4829	877	39	Male	Professional	High School	Single	4	61284.85	36604.93	1214.27	2023-12-15	19:28:28
12301	4394	39	Male	Student	Bachelor	Single	2	28419.25	11172.93	600.92	2023-02-28	01:51:57
9933	103	21	Male	Student	High School	Single	3	57384.51	40721.71	10819.52	2023-09-10	10:57:08
10840	3288	52	Female	Professional	Master	Single	1	40702.41	21169.11	63.55	2023-10-24	16:28:04
8798	358	56	Male	Student	Bachelor	Widowed	3	58628.6	17873.88	1082.43	2023-09-29	22:29:05
16671	1732	41	Male	Student	Bachelor	Married	2	85310.21	80911.26	4847.91	2023-12-24	20:08:05
4605	3336	31	Male	Professional	High School	Single	0	93197.13	68415.46	3304.34	2023-12-12	21:06:00
8238	4228	35	Male	Retired	Bachelor	Single	3	118437.1	39882.13	2740.89	2023-10-30	19:38:10
10643	2006	41	Female	Professional	Bachelor	Single	2	68140.84	21877.39	2370.43	2023-04-15	05:14:44
1835	65	21	Female	Professional	Bachelor	Divorced	1	53705.83	41566.98	1951.75	2023-10-28	05:42:44
17034	1116	32	Female	Professional	High School	Married	1	60108.48	60993.62	2331.67	2023-04-12	01:46:26
8097	4530	18	Male	Unemployed	Bachelor	Widowed	0	58974.57	33755.03	333.98	2023-12-23	16:43:20
15195	4143	30	Female	Professional	High School	Single	4	127088.35	131498.77	15901.04	2023-12-15	02:29:35
8237	1162	29	Female	Student	Bachelor	Married	1	42903.25	31294.62	2936.4	2023-12-02	11:18:44
16688	1737	36	Male	Professional	High School	Married	1	68350.31	60727.89	873.61	2023-09-24	00:24:13
17924	382	35	Female	Student	High School	Single	0	78404.02	99725.62	6948.17	2023-10-08	08:53:21
5852	2587	28	Female	Professional	Bachelor	Single	1	66233.88	23587.19	9162.02	2023-09-25	23:21:12
11620	3113	31	Female	Professional	High School	Married	2	127411.46	54214.57	3895.3	2023-11-20	21:55:22
9616	1371	30	Male	Professional	High School	Widowed	4	53477.15	30523.44	2239.89	2023-11-21	11:01:45
18255	1066	28	Female	Retired	Bachelor	Single	1	76889.65	69676.4	5576.9	2023-09-06	15:59:49

Overview of Model Functionality

- Gradient boosters are widely regarded as the best model for classification tasks with numerical data.
- Also, alternative approaches were tried, like a random forest classifier and a neural network; however, they yielded unsatisfactory results of around 50%.
- Not that many rows had missing data, and those were deleted. A lot of the data, however, was in the wrong format or had unreasonable values. The wrongly formatted data was cleaned and standardised; unreasonable data was deleted.
- For training, categorical variables were encoded using the one-hot encoding technique. This technique is used on non-ordinal categories, which are the ones that were used for training.
- A part of the data was reserved for testing and was not used for training to accurately assess model performance. Metrics like accuracy and F1-score were used.